# Ontology Summarization Using Labelled Latent Dirichlet Allocation Method

**Dr .A.Mekala**
Assistant Professor
Department of Computer Application
Sacred Heart College , Tamilnadu , India.

***Abstract: With the advent of the Internet, the amount of Semantic Web documents that describe real-world entities and their inter-links as a set of statements have grown considerably. These descriptions are usually lengthy, which makes the utilization of the underlying entities a difficult task. Entity summarization, which aims to create summaries for real world entities, has gained increasing attention in recent years. we present hierarchical relation based Labelled Latent Dirichlet Allocation (L-LDA), a data-driven hierarchical topic model for extracting terminological ontologies from a large number of heterogeneous documents. In contrast to traditional topic models, L-LDA relies on noun phrases instead of unigrams, considers syntax and document structures, and enriches topic hierarchies with topic relations. Through a series of experiments, we demonstrate the superiority of L-LDA over existing topic models, especially for building hierarchies. Furthermore, we illustrate the robustness of L-LDA in the settings of noisy data sets, which are likely to occur in many practical scenarios. Our ontology evaluation results show that ontologies extracted from L-LDA are very competitive with the ontologies created by domain experts.***

***Keywords: Ontology, L-LDA, Recall, Precision, F-Measure Rate.***

---

## 1. Introduction

Topic models such as Labelled Latent Dirichlet Allocation (L-LDA) have gained considerable attention, recently. They have been successfully applied to a wide variety of text mining tasks, such as word sense disambiguation, sentiment analysis, information retrieval and others, in order to identify hidden topics in text documents. Topic models typically assume that documents are mixtures of topics, while topics are probability distributions over the vocabulary. When the topic proportions of documents are estimated, they can be used as the themes (high level representations of the semantics) of the documents. Highest-ranked words in a topic-word distribution indicate the meaning of the topic. Thus, topic models provide an effective framework for extracting the latent semantics from unstructured text collections.

Automatic text summarization is the task of producing a concise and fluent summary while preserving key information content and overall meaning. In recent years, numerous approaches have been developed for automatic text summarization and applied widely in various domains. For example, search engines generate snippets as the previews of the documents. Other examples include news websites which produce condensed descriptions of news topics usually as headlines to facilitate browsing or knowledge extractive approaches. Automatic text summarization is very challenging, because when we as humans summarize a piece of text, we usually read it entirely to develop our understanding, and then write a summary highlighting its main points. Since computers lack human knowledge and language capability, it makes automatic text summarization a very difficult and non-trivial task. We propose an ontology-based topic model, L-LDA, which incorporates an ontology into the topic model in a systematic manner. Our model integrates the topics to external knowledge bases, which can benefit other research areas such as information retrieval, classification and visualization. We introduce a topic labeling method, based on the semantics of the concepts that are included in the discovered topics, as well as ontological relationships existing among the concepts in the ontology. Our model improves the labeling accuracy by exploiting the topic-concept relations and can automatically generate labels that are meaningful for interpreting the topics. We demonstrate the usefulness of our approach in two ways. We first show how our model can be exploited to link text documents to ontology concepts and categories. Then we illustrate automatic topic labeling by performing a series of experiments.

## 2. Literature Survey

**Mei et al.** proposed an approach to automatically label the topics by converting the labeling problem to an optimization problem. First they generate candidate labels by extracting either bigrams or noun chunks from the collection of documents. Then, they rank the candidate labels based on Kullback-Leibler (KL) divergence with a given topic, and choose a candidate label that has the minimum KL divergence and the maximum mutual information with the topic to label the corresponding topic. Aan algorithm for topic labeling based on a given topic hierarchy. Given a topic, they generate label candidate set using Google Directory hierarchy and find the best label according to a set of similarity measures. **Lau et al.** introduced a method for topic labeling by selecting the best topic word as its label based on a number of features. They assume that the topic terms are representative enough and appropriate to be considered as labels, which is not always the case. Reused the features proposed and also extended the set of candidate labels exploiting Wikipedia. For each topic they first select the top terms and query the Wikipedia to find top article titles having the these terms according to the features and consider them as extra candidate labels. Then they rank the candidate to find the best label for the topic. **Mao et al.** proposed a topic labeling approach which enhances the labeling by using the sibling and parent-child relations between topics. They first generate a set of candidate labels by extracting meaningful phrases using Ngram Testing for a topic and adding the top topic terms to the set based on marginal term probabilities. And then rank the candidate labels by exploiting the hierarchical structure between topics and pick the best candidate as the label of the topic. **Hulpus et al.** proposed an automatic topic labeling approach by exploiting structured data from DBpedia2 . Given a topic, they first find the terms with highest marginal probabilities, and then determine a set of DBpedia concepts where each concept represents the identified sense of one of the top terms of the topic. After that, they create a graph out of the concepts and use graph centrality algorithms to identify the most representative concepts for the topic. **Mimno et al.**

proposed the hPAM model that models a document as a mixture of distributions over super-topics and sub-topics, using a directed acyclic graph to represent a topic hierarchy. The OntoLDA model is different, because in hPAM, distribution of each super-topic over sub-topics depends on the document, whereas in OntoLDA, distributions of topics over concepts are independent of the corpus and are based on an ontology. The other difference is that sub-topics in the hPAM model are still unigram words, whereas in OntoLDA, ontological concepts are n-grams, which makes them more specific and more meaningful, a key point in OntoLDA.

**3. Proposed Work**

The Labelled Latent Dirichlet Allocation (L-LDA) is a generative probabilistic model for extracting thematic information (topics) from a collection of documents. L-LDA assumes that each document is made up of various topics, where each topic is a probability distribution over words. Let d = {$d_1$, $d_2$, . . . $d_{|D|}$}be a corpus of documents and V = ($w_1$, w2, . . $w_{|v|}$) a vocabulary of the corpus. A topic $z_j$, 1 <= j <= K is represented as a multinomial probability distribution over the |V| words, p($w_i$, $z_i$).

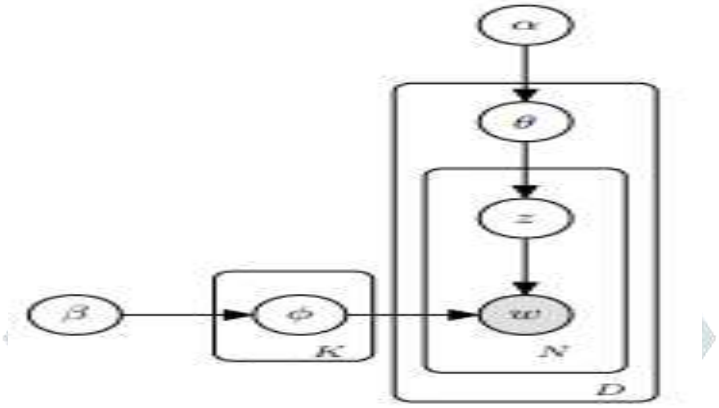


**Figure 1: L-LDA Graphical Model**

The L-LDA model is the typical representative of topic models. L-LDA is a generative probabilistic model for collections of discrete data such as text corpora. Documents are represented as random mixtures over latent topics, and each topic is then characterized by a distribution over words, shown in Figure 1. The text generative model describes the generative process of words through documents based on latent variable and simple probabilistic sampling rules, and probabilistic topic models have been used to analyze the topic structure of given texts and the implication of each word.

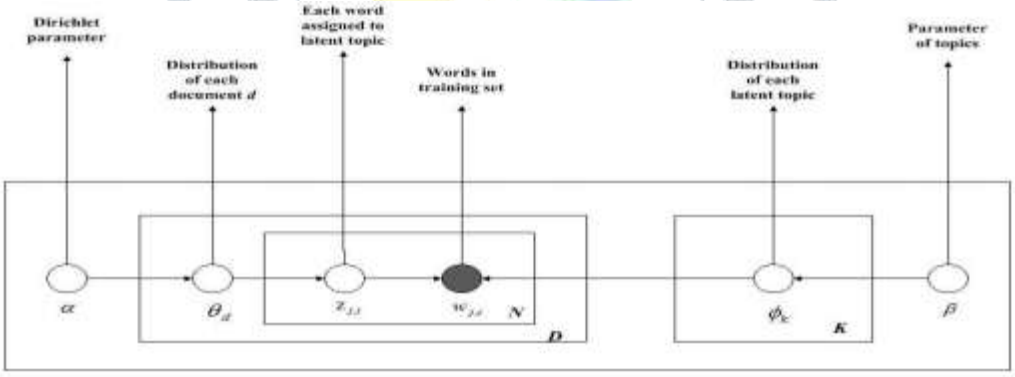


**Figure 2: Basic Idea of L-LDA Model**

The L-LDA model is also considered a bag-of-words model with three levels displayed in Figure 2. As the figure clearly shows, the corpus contains a collection of D documents. Each document w consists of K topics, and each topic k is characterized by a distribution over N words. L-LDA is a probabilistic generative model for modeling entities in RDF graphs. The key idea behind our model is twofold: we exploit statistical topic models as the underlying quantitative framework for entity summarization; and L-LDA incorporates the prior knowledge from the RDF knowledge base directly into the topic model.

**3.1 L-LDA Modeling Module**

The basic principle of the L-LDA model has already been introduced, and the specific steps are as follows: First, initialize the pre-processing text, thus, match the text with the model matrix in the topic database in order to select the most suitable model entering the training set; Afterward, the training set will generate characteristic sequences, which is the data source for L-LDA modeling. Then, according to the basic principle of the L-LDA model, the characteristic sequence will be modeled to generate a topic-word model matrix. Finally, output and save the matrix to the topic database in order to prepare for the next round.

**4. Experimental Results**

Labeled LDA is defined as a probabilistic graphical model that illustrates a process for generating a labeled collection of documents. Similar to LDA, Labeled Latent Dirichlet Allocation models a document as a combination of original topics and generates each word from one topic. In contradiction to L-LDA constrains the semantically obtained topic model from refined ontology to use only those topics that relates to a document's (observed) label set.

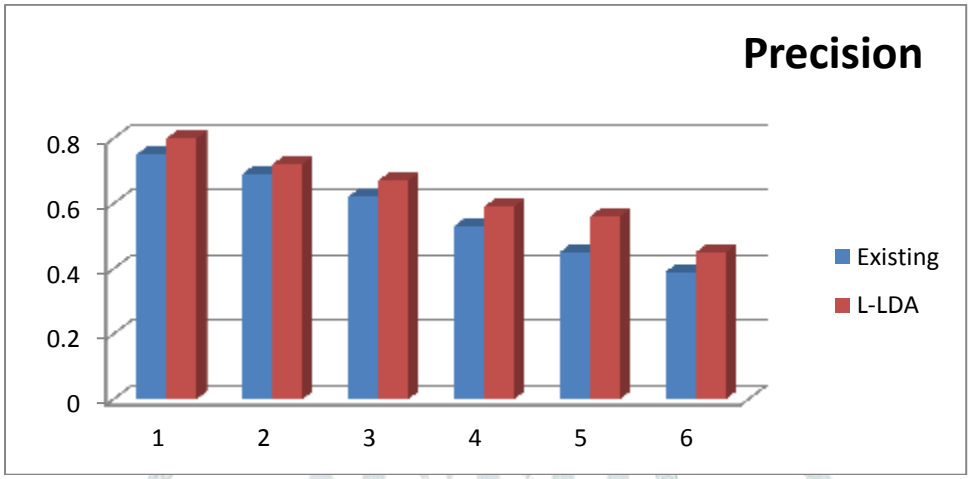


**Figure 3: Precision**

Figure 3 represents precision values are compare with existing and proposed L-LDA values. L-LDA values are higher than compare with existing values.

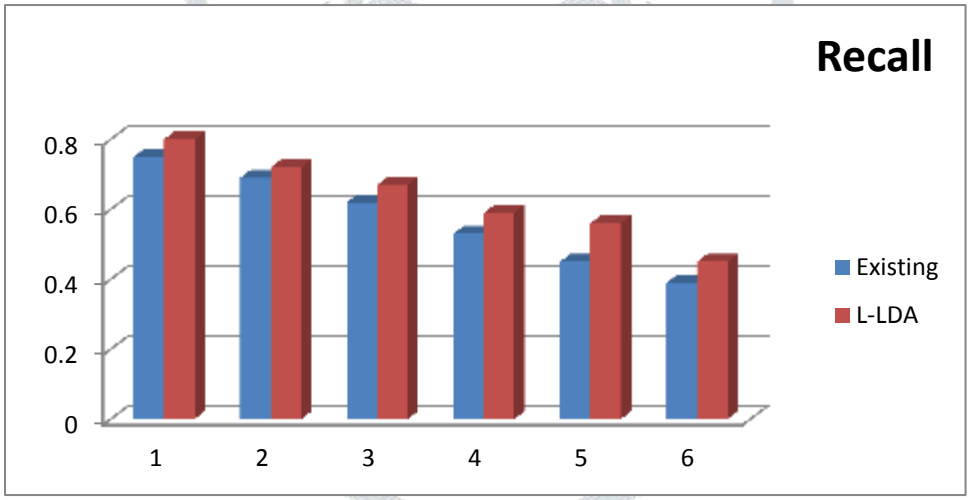


**Figure 4: Recall**

Figure 4 represents recall values are compare with existing and proposed L-LDA values. L-LDA values are higher than compare with existing values.

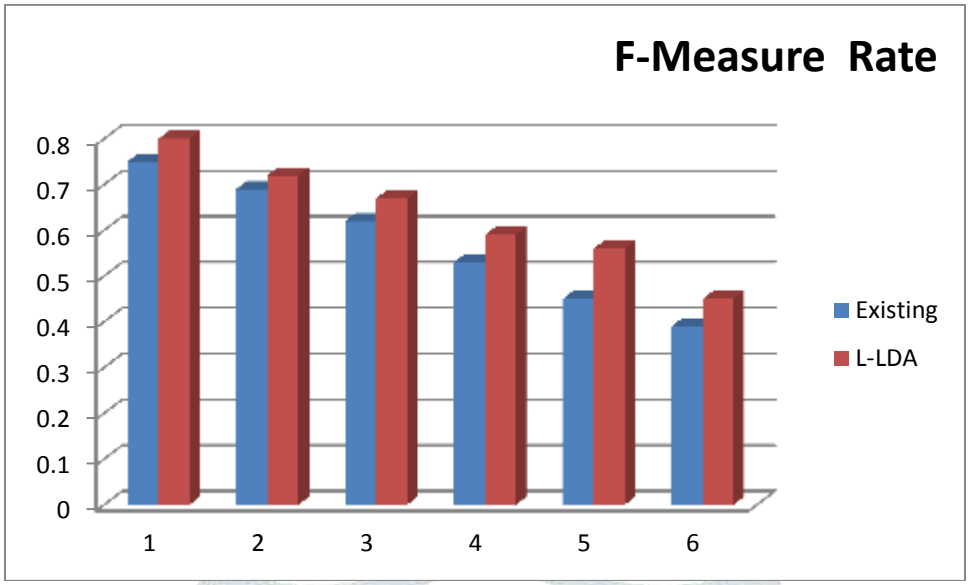


**Figure 5: F-Measure Rate**

Figure 5 represents F-Measure Rate values are compare with existing and proposed L-LDA values. L-LDA values are higher than compare with existing values.

**Conclusion**

L-LDA, an ontology based topic model, along with a graph-based topic labeling method for the task of topic labeling. Experimental results show the effectiveness and robustness of the proposed method when applied on different domains of text collections. The proposed ontology based topic model improves the topic coherence in comparison to the standard L-LDA model by integrating ontological concepts with probabilistic topic models into a unified framework. There are many interesting future research directions of this work. It would be interesting to investigate how this model and a much richer set of topic models that combine prior knowledge with statistical learning techniques could be used for various tasks in the Semantic Web domain, such as ontology summarization, ontology tagging, and finding similar ontologies.

**References:**

[1] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene. Unsupervised graph-based topic labelling using dbpedia. In Proceedings of the sixth ACM international conference on Web search and data mining, pages 465–474. ACM, 2013.
[2] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 46(5):604–632, 1999.
[3] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin. Automatic labelling of topic models. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 1536–1545. Association for Computational Linguistics, 2011.
[4] J. H. Lau, D. Newman, S. Karimi, and T. Baldwin. Best topic word selection for topic labelling. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 605–613. Association for Computational Linguistics, 2010.
[5] A. Lazaridou, I. Titov, and C. Sporleder. A bayesian model for joint unsupervised induction of sentiment, aspect and discourse representations. In ACL (1), pages 1630–1639, 2013.
[6] C. Li, A. Sun, and A. Datta. A generalized method for word sense disambiguation based on wikipedia. In Advances in Information Retrieval, pages 653–664. Springer, 2011.
[7] C. Li, A. Sun, and A. Datta. Tsdw: Two-stage word sense disambiguation using wikipedia. Journal of the American Society for Information Science and Technology, 64(6):1203–1223, 2013.
[8] J. Li, C. Cardie, and S. Li. Topicspam: a topic-model based approach for spam detection. In ACL (2), pages 217–221, 2013.
[9] C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. In Proceedings of the 18th ACM conference on Information and knowledge management, pages 375–384. ACM, 2009.
[10] Q. Luo, E. Chen, and H. Xiong. A semantic term weighting scheme for text categorization. Expert Systems with Applications, 38(10):12708–12716, 2011.
[11] D. Magatti, S. Calegari, D. Ciucci, and F. Stella. Automatic labeling of topics. In Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on, pages 1227–1232. IEEE, 2009.
[12] X.-L. Mao, Z.-Y. Ming, Z.-J. Zha, T.-S. Chua, H. Yan, and X. Li. Automatic labeling hierarchical topics. In Proceedings of the 21st ACM international conference on Information and knowledge management, pages 2383–2386. ACM, 2012.
[13] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In Proceedings of the 15th international conference on World Wide Web, pages 533–542. ACM, 2006.
[14] Q. Mei, X. Shen, and C. Zhai. Automatic labeling of multinomial topic models. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 490–499. ACM, 2007.